

Sample Final Exam 1 – Part A

Question 1

Let X_1 be the student's score on the first midterm, let X_2 be the score on the second midterm and let X_3 be the score on the final exam. We calculate the student's final percentage grade as a weighted average:

$$0.2(70) + 0.3(60) + 0.5x_3 \geq 75 \quad \Rightarrow \quad 32 + 0.5x_3 \geq 75$$

$$\Rightarrow \quad 0.5x_3 \geq 43 \quad \Rightarrow \quad x_3 \geq \frac{43}{0.5} = 86$$

The student needs a minimum of 86 on the final exam to get a B⁺ in the course.

Question 2

We are using a child's milk consumption to explain/predict his or her height, so milk consumption is the explanatory variable X and height is the response variable Y . We calculate the slope and intercept of the least squares regression line:

$$b_1 = r \frac{s_y}{s_x} = 0.4 \left(\frac{3.0}{1.5} \right) = 0.8, \quad b_0 = \bar{y} - b_1 \bar{x} = 60.0 - 0.8(4.0) = 56.8$$

The equation of the least squares regression line is $\hat{y} = 56.8 + 0.8x$. The predicted height of a student who drinks 5 cups of milk per day is

$$\hat{y} = 56.8 + 0.8(5) = 60.8$$

Question 3

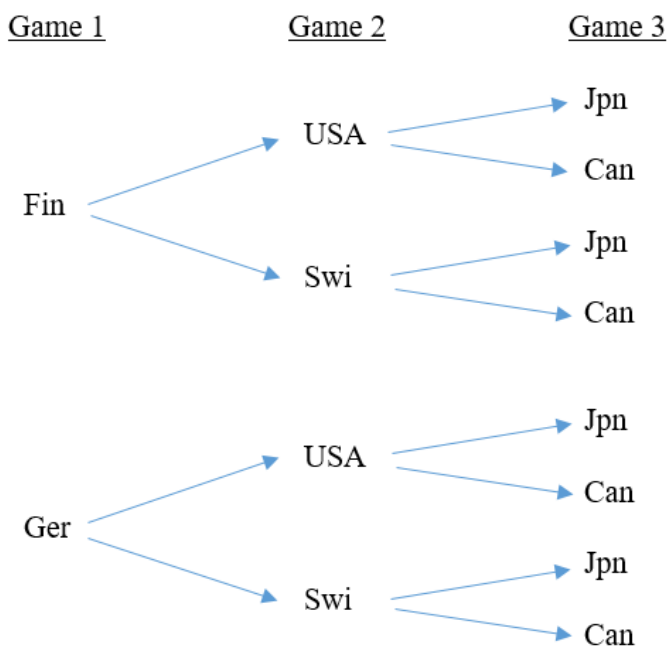
A separate simple random sample is selected from each of three strata (the lakes). The combined sample is a stratified sample.

Question 4

The teachers and times were not randomly assigned to the students (nothing was done randomly). This is an observational study, not an experiment.

Question 5

To determine the total number of outcomes, we could draw a tree diagram, but since there are five games, it would be quite large and time consuming. Suppose only the first three games were being played. Let's look at the tree diagram for this situation:



The total number of outcomes would be $2 \times 2 \times 2 = 8$. If we add the fourth game, we draw two arrows from each of these 8 outcomes, and we get 16 outcomes. If we add a fifth game, we draw two more arrows from each of these 16 outcomes, and we get 32 outcomes. The short answer is that the number of outcomes is always equal to 2 to the power of the number of games, in this case $2^5 = 32$.

Question 6

There are many different ways for at least one underdog to win. There is only ONE way for this NOT to happen. Therefore,

$$\begin{aligned} P(\text{at least one underdog wins}) &= 1 - P(\text{no underdogs win}) = 1 - P(\text{all favourites win}) \\ &= 1 - P(Ger \cap Swi \cap Can \cap Swe \cap Sco) = 1 - P(Ger)P(Swi)P(Can)P(Swe)P(Sco) \\ &= 1 - (0.57)(0.72)(0.89)(0.67)(0.82) = 1 - 0.20 = 0.80 \end{aligned}$$

Note that we can multiply probabilities for the five games, since the winner of one game is independent of the winner of any other game.

Question 7

We know

$$P(B \cup V) = P(B) + P(V) - P(B \cap V)$$

$$\Rightarrow P(V) = P(B \cup V) - P(B) + P(B \cap V)$$

$$\Rightarrow P(V) = 0.50 - 0.40 + 0.05 = 0.15$$

Question 8

We know that none of the three pairs of events are mutually exclusive, since $P(W \cap B) > 0$, $P(W \cap V) > 0$ and $P(B \cap V) > 0$. (If two of the events were mutually exclusive, it would mean it is impossible for a person to buy both types of alcohol, which obviously isn't true.)

In order to determine if two events A and B are independent, we must check if $P(A \cap B) = P(A)P(B)$. We check for each pair of events:

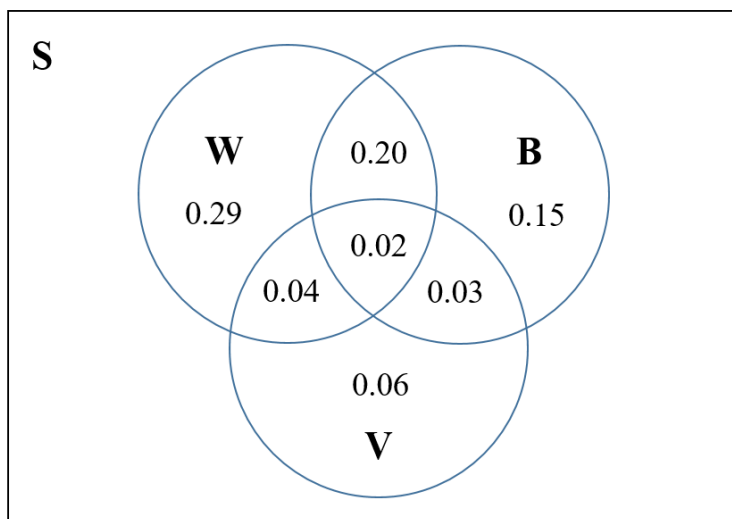
$$P(W)P(B) = (0.55)(0.40) = 0.22 = P(W \cap B), \text{ so } W \text{ and } B \text{ **are** independent.}$$

$$P(W)P(V) = (0.55)(0.15) = 0.0825 \neq P(W \cap V) = 0.06, \text{ so } W \text{ and } V \text{ are **not** independent.}$$

$$P(B)P(V) = (0.40)(0.15) = 0.06 \neq P(B \cap V) = 0.05, \text{ so } B \text{ and } V \text{ are **not** independent.}$$

Question 9

We draw a Venn diagram, finding the probability of each mutually exclusive part (starting from the inside with $P(W \cap B \cap V)$ and working our way out). The Venn diagram is shown below:



We see the probability that a customer buys only wine is $P(W \cap B^c \cap V^c) = 0.29$. Note that we could have calculated this probability as follows:

$$\begin{aligned} P(W \cap B^c \cap V^c) &= P(W) - P(W \cap B) - P(W \cap V) + P(W \cap B \cap V) \\ &= 0.55 - 0.22 - 0.06 + 0.02 = 0.29 \end{aligned}$$

(We had to add $P(W \cap B \cap V)$ at the end, because we subtracted it twice, as it was included in both $P(W \cap B)$ and $P(W \cap V)$).

Note that when you decompose a Venn diagram like this, you can find any probability just by adding the appropriate parts. For example, the probability that someone buys exactly two types of alcohol is

$$P(W \cap B \cap V^c) + P(W \cap B^c \cap V) + P(W^c \cap B \cap V) = 0.20 + 0.04 + 0.03 = 0.27$$

The probability that someone buys wine but not beer is $0.29 + 0.04 = 0.33$.

The probability that someone doesn't buy any of the three types of alcohol is one minus the total of the probabilities inside the three circles:

$$P(W^c \cap B^c \cap V^c) = 1 - (0.29 + 0.20 + 0.15 + 0.04 + 0.02 + 0.03 + 0.06) = 1 - 0.79 = 0.21$$

Question 10

Let X be the number of the 17 customers that buy wine. There is a fixed sample size ($n = 17$), there are only two outcomes of interest (a person buys wine or they don't), outcomes are independent and there is a constant probability of success ($p = 0.55$). As such, X has a binomial distribution with parameters $n = 17$ and $p = 0.55$, and so

$$P(X = 10) = \binom{17}{10} (0.55)^{10} (0.45)^7 = 0.1841$$

Question 11

Let X be the number of the 300 customers that buy beer. Then X has a binomial distribution with parameters $n = 300$ and $p = 0.40$. Since the sample size is very large, we can use the normal approximation (you can check that np and $n(1 - p)$ are both greater than 10):

$$\begin{aligned} P(\hat{p} < 0.38) &\approx P\left(Z < \frac{0.38 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(Z < \frac{0.38 - 0.40}{\sqrt{\frac{0.40(0.60)}{300}}}\right) \\ &= P(Z < -0.71) = 0.2389 \end{aligned}$$

Question 12

$$\begin{aligned} P(\text{voted for different parties}) &= 1 - P(\text{voted for same party}) \\ &= 1 - (P(NN) + P(LL) + P(CC)) = 1 - (0.40(0.40) + 0.35(0.35) + 0.25(0.25)) \\ &= 1 - 0.345 = 0.655 \end{aligned}$$

Note that we can multiply probabilities for the two voters, since the party one person voted for is independent of any other person.

Question 13

Let X be the number of patients who require attention at any given time. There is a fixed number of individuals ($n = 4$), there are only two outcomes of interest (a patient needs attention or she doesn't), whether one patient needs attention is independent of any other, and there is a constant probability of needing attention ($p = 0.3$). The probability that there aren't enough nurses for everyone who needs attention, i.e., the probability that at least 3 patients need attention, is:

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) = \binom{4}{3}(0.3)^3(0.7)^1 + \binom{4}{4}(0.3)^4(0.7)^0 \\ &= 0.0756 + 0.0081 = 0.0837 \end{aligned}$$

Question 14

It helps to draw the normal curve for this question and shade in the appropriate area. We know that $P(X > 1200) = 0.2000$.

We first find the value z such that $P(Z > z) = 0.2000 \Rightarrow P(Z \leq z) = 0.8000$. We find this value to be $z = 0.84$. We now use the z -score formula to solve for μ :

$$z = \frac{x - \mu}{\sigma} \Rightarrow \mu = x - z\sigma = 1200 - 0.84(172) = 1055.52 \approx 1055.5$$

Question 15

The first quartile is the value x such that $P(X < x) = 0.2500$. We first find the value z such that $P(Z < z) = 0.2500$ to be $z = -0.67$. We now use the z -score formula to solve for μ :

$$z = \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma = 90 + (-0.67)(20) = 76.6$$

Question 16

Since X has a normal distribution, $\bar{X}$ has an exact normal distribution, even though n is low. We calculate the probability as follows:

$$\begin{aligned} P(80 < \bar{X} < 100) &= P\left(\frac{80 - 90}{20/\sqrt{3}} < Z < \frac{100 - 90}{20/\sqrt{3}}\right) = P(-0.87 < Z < 0.87) \\ &= P(Z < 0.87) - P(Z < -0.87) = 0.8078 - 0.1922 = 0.6156 \end{aligned}$$

Question 17

The variable X doesn't follow a normal distribution, and the sample size is too low for the Central Limit Theorem to apply. Since the distribution of $\bar{X}$ is unknown (not normal), we cannot calculate this probability.

Question 18

Although the population is not normally distributed, the sample size is high, so according to the Central Limit Theorem, $\bar{X}$ is approximately normally distributed. Therefore,

$$P(\text{Total} < 1000) = P\left(\bar{X} < \frac{1000}{40}\right) = P(\bar{X} < 25)$$

$$\approx P\left(Z < \frac{25 - 24}{15/\sqrt{40}}\right) = P(Z < 0.42) = 0.6628$$

Question 19

For a fixed confidence level, when we **decrease** the margin of error by a factor of k , the required sample size is k^2 times **greater**. Here, $k = 60/20 = 3$, i.e., we are decreasing the margin of error by a factor of 3. Therefore, we need a sample size $3^2 = 9$ times greater, i.e., $n = 9(90) = 810$. This can be seen by looking at the sample size formula:

$$n = \left(\frac{z^*\sigma}{m}\right)^2$$

When the margin of error m becomes 3 times smaller, the fraction becomes 3 times **larger**. Then we square the quantity inside the parentheses, so the required sample size is $3^2 = 9$ times larger.

Question 20

The formula for the confidence interval is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

In order to determine the confidence level, we need to find the critical value z^* . We know that $\sigma = 2.4$ and $n = 4$. We also know that $\bar{x}$ is at the centre of the confidence interval, so

$$\bar{x} = \frac{128.609 + 134.191}{2} = 131.4$$

We can find z^* using either the lower or upper limit of the confidence interval. We will use the upper limit:

$$\bar{x} + z^* \frac{\sigma}{\sqrt{n}} = 134.191$$

$$\Rightarrow 131.4 + z^* \left(\frac{2.4}{\sqrt{4}} \right) = 134.191$$

$$\Rightarrow z^* = (134.191 - 131.4) \left(\frac{\sqrt{4}}{2.4} \right) = 2.326$$

From the bottom of Table 2, we see that this critical value corresponds to a 98% confidence interval.

Question 21

Even though the population isn't normally distributed, the Central Limit Theorem tells us that, since n is large, the sampling distribution of $\bar{X}$ is approximately normal with mean $\mu = 0$ and standard deviation $\sigma/\sqrt{n} = 0.4/\sqrt{100} = 0.04$.

Question 22

We are testing $H_0: \mu = 400$ vs. $H_a: \mu > 400$.

$$\text{The test statistic is } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{430 - 400}{165/\sqrt{50}} = 1.29.$$

The P-value is $P(Z \geq 1.29) = 1 - P(Z < 1.29) = 1 - 0.9015 = 0.0985$.

Question 23

Since the P-value $> \alpha$, we fail to reject H_0 . We have insufficient evidence that H_a is true, i.e., insufficient evidence that the man is lying. (Note that we do **not** conclude that the man is telling the truth. We don't have enough evidence that he is lying, but that doesn't mean he's telling the truth.)

Question 24

We are conducting an upper-tailed test at $\alpha = 0.05$. We reject H_0 , so we know the P-value $< \alpha = 0.05$.

Consider Statement (I). We reject H_0 if the P-value $< \alpha$. We know the P-value is less than 0.05, since we rejected H_0 at $\alpha = 0.05$, but we don't know if the P-value is less than 0.01. Therefore, we don't know if we would reject H_0 at $\alpha = 0.10$. Statement (I) is not necessarily true.

Note that if we were given the information the other way around – i.e., if we rejected H_0 at a 1% level of significance, we would know the P-value is less than 0.01, so it would obviously be less than 0.05. Therefore, we would definitely reject H_0 at the 5% level of significance.

Consider Statement (II). The sample mean that was calculated in the test was high enough to reject H_0 and conclude that $\mu > 10$. Then it is definitely high enough to conclude that $\mu > 9$. If the sample mean we got is strong enough evidence to conclude μ is greater than 10, then it must be enough to conclude that μ is greater than 9. Statement (II) is true. (Note that the converse would not necessarily be true – if we calculated an $\bar{x}$ that was high enough to conclude that $\mu > 9$, that does NOT necessarily mean it's high enough to conclude that $\mu > 10$.)

Consider Statement (III). We know the P-value for a two-sided test is DOUBLE the P-value for a one-sided test. The P-value of the original one-sided test is less than 0.05. The P-value in (III) would be double this value. Would it still be less than 0.05? We don't know! In other words, we don't know for sure whether we would reject H_0 if this was a two-sided test. Statement (III) is not necessarily true.

Note that if we were given the information the other way around – i.e., if we rejected H_0 for the two-sided test – then we would definitely reject H_0 for the one-sided test, since the P-value for the one-sided test is only half the P-value for the two-sided test. It would definitely still be below 0.05.

Question 25

We are testing $H_0: \mu = 150$ vs. $H_a: \mu \neq 150$. Since we don't know the value of the population standard deviation σ , the test statistic is t .

The test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{162 - 150}{35/\sqrt{30}} = 1.88$.

The test statistic follows a t distribution with $n - 1 = 29$ degrees of freedom. The P-value is $2P(T(29) \geq 1.88)$. From Table 2 that

$$P(T(29) \geq 1.699) = 0.05 \quad \text{and} \quad P(T(29) \geq 2.045) = 0.025$$

Since $1.699 < t = 1.88 < 2.045$, it follows that $P(T(29) \geq 1.88)$ is between 0.025 and 0.05. But our P-value is double this probability, so the P-value is between 0.05 and 0.10.

Since the P-value $> \alpha = 0.01$, we fail to reject H_0 .

Question 26

We are conducting a matched pairs t test. We are testing the hypotheses $H_0: \mu_d = 0$ vs. $H_a: \mu_d < 0$. Note that this is a lower-tailed test because we wanted to know whether the mean time is longer AFTER consuming the wine. The differences are calculated as $d = S - AW$, so we must ask, "If times are longer on average after drinking wine, would we expect the differences to be positive or negative?". In fact, we would expect them to be negative, as we would expect times while sober to be LOWER, and thus $d = S - AW$ should be negative.

Question 27

The test statistic is $t = \frac{\bar{x}_d}{s_d/\sqrt{n}} = \frac{-29.4}{23.5/\sqrt{5}} = -2.80$.

Under H_0 , this test statistic follows a t distribution with $n - 1 = 5 - 1 = 4$ degrees of freedom. The P-value is $P(T(4) \leq -2.80) = P(T(4) \geq 2.80)$ by the symmetry of the t distributions. From Table 2, we see that

$$P(T(4) \geq 2.776) = 0.025 \quad \text{and} \quad P(T(4) \geq 2.999) = 0.02$$

Since $2.776 < t = 2.80 < 2.999$, the P-value is between 0.02 and 0.025.

Question 28

We need to find the critical value z^* for a 92% confidence interval. Draw the standard normal curve for Z , centred at $\mu = 0$. Now draw lines at the values $-z^*$ and z^* on either side of the mean (equally far in either direction), such that the area between $-z^*$ and z^* is 0.92. In order to find the value z^* , we need to find the total area to the left of z^* (so we can use the table to look it up). From the picture, you see that this is equal to the area to the left of $-z^*$ plus the area between $-z^*$ and z^* . We already know $P(-z^* < Z < z^*) = 0.9200$, so we need to find $P(Z < -z^*)$. We know the central area (between $-z^*$ and z^*) is 0.9200, so the total area of the two tails (i.e., below $-z^*$ and above z^*) is $1 - 0.9200 = 0.0800$. That means each tail has an area equal to half this value, $0.0800/2 = 0.0400$. Therefore, the total area to the left of z^* is

$$P(Z < z^*) = P(Z < -z^*) + P(-z^* < Z < z^*) = 0.0400 + 0.9200 = 0.9600.$$

Using the normal table (going backwards), we find that the value of z^* with 0.9600 area to its left is $z^* = 1.75$. (We take the closest value in the table.)

The required sample size is $n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*) = \left(\frac{1.75}{0.035}\right)^2 (0.5)(0.5) = 625$.

In sample size calculations for proportions, unless the question specifically states some other value of p , we always use $p^* = 0.5$ in our calculation. This gives us the maximum sample size we could possibly need to achieve our margin of error, regardless of the true value of p .

Question 29

The formula for the confidence interval for p is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$.

where the quantity after the “ $\pm$ ” is the margin of error. In this case, $\hat{p} = 42/200 = 0.21$, and the margin of error is

$$m = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 1.96 \sqrt{\frac{0.21(0.79)}{200}} = 0.0565$$

Question 30

We are testing $H_0: p = 0.08$ vs. $H_a: p < 0.08$. We calculate $\hat{p} = 15/300 = 0.05$.

The test statistic is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.05 - 0.08}{\sqrt{\frac{0.08(0.92)}{300}}} = -1.92$.

The P-value is $P(Z \leq -1.92) = 0.0274$.

Sample Final Exam 1 – Part B

1. (a) The minimum is 2.2 and the maximum is 11.4.

The median is in position $(30 + 1)/2 = 15.5$, i.e., the average of the 15th and 16th ordered data values. Therefore, the median is equal to $(7.0 + 7.2)/2 = 7.1$.

The first quartile is the median of all data values lower in position than the median, so $Q1$ is in position $(15 + 1)/2 = 8$. Therefore, $Q1 = 6.0$.

The third quartile is the median of all data values higher in position than the median, so $Q3$ is 8 positions above the median (or equivalently, 8 positions down from the maximum). Therefore, $Q3 = 8.0$.

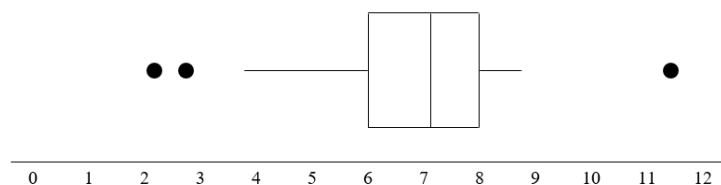
The five-number summary is: 2.2 6.0 7.1 8.0 11.2

- (b) We calculate the fences:

$$LF = Q1 - 1.5IQR = 6.0 - 1.5(8.0 - 6.0) = 3.0$$

$$UF = Q3 + 1.5IQR = 8.0 + 1.5(8.0 - 6.0) = 11.0$$

There are two values below the lower fence (2.2 and 2.8) and one value above the upper fence (11.4). These three values will be labelled as outliers. The whisker on the left side of the box extends to the lowest value that is not an outlier, which is 3.8 (i.e., the “new minimum”). The whisker on the right side of the box extends to the highest value that is not an outlier, which is 8.8 (i.e., the “new maximum”). The outlier boxplot is shown below:



Excluding outliers, the distribution is skewed to the left.

2. (a) When the age of a car increases by one year, we predicted its price to decrease by \$1,700.
- (b) We are told that 81% of the variation in a car's price can be accounted for by its regression on age. This is the definition of r^2 . We know $r^2 = 0.81$, so the correlation is $r = -\sqrt{0.81} = -0.90$. Note that we take the negative square root, as we see from the slope of the regression line that the relationship between age and price is negative.
- (c) The residual for this car is

$$y_i - \hat{y}_i = 15500 - (18600 - 1700(5)) = 15500 - 10100 = 5400$$

3. (a) The sample space of outcomes is shown below, as well as the probability of each outcome:

Outcome	Probability
GGG	$(0.2)(0.4)(0.3) = 0.024$
GGC	$(0.2)(0.4)(0.7) = 0.056$
GSG	$(0.2)(0.6)(0.3) = 0.036$
GSC	$(0.2)(0.6)(0.7) = 0.084$
SGG	$(0.5)(0.4)(0.3) = 0.060$
SGC	$(0.5)(0.4)(0.7) = 0.140$
SSG	$(0.5)(0.6)(0.3) = 0.090$
SSC	$(0.5)(0.6)(0.7) = 0.210$
CGG	$(0.3)(0.4)(0.3) = 0.036$
CGC	$(0.3)(0.4)(0.7) = 0.084$
CSG	$(0.3)(0.6)(0.3) = 0.054$
CSC	$(0.3)(0.6)(0.7) = 0.126$

Note that we multiply probabilities, since the colour coin chosen from one hat is independent of any other.

- (b) The variable X can take the values 0, 1, 2 or 3. We find the probability distribution of X as follows:

$$P(X = 0) = P(SSC) + P(CSC) = 0.210 + 0.126 = 0.336$$

$$\begin{aligned} P(X = 1) &= P(GSC) + P(SGC) + P(SSG) + P(CGC) + P(CSG) \\ &= 0.084 + 0.140 + 0.090 + 0.084 + 0.054 = 0.452 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= P(GGC) + P(GSG) + P(SGG) + P(CGG) \\ &= 0.056 + 0.036 + 0.060 + 0.036 = 0.188 \end{aligned}$$

$$P(X = 3) = 0.024$$

The probability distribution of X is summarized in the table below:

x	0	1	2	3
$P(X = x)$	0.336	0.452	0.188	0.024

4. The Central Limit Theorem states that, regardless of the population distribution of a variable X , the sampling distribution of $\bar{X}$ will be approximately normal, provided that the sample size is high (we use $n \geq 30$).
5. (a) A 98% confidence interval for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 28.4 \pm 2.479 \left(\frac{5.3}{\sqrt{27}} \right) = 28.4 \pm 2.53 = (25.87, 30.93)$$

where $t^* = 2.479$ is the upper 0.01 critical value for the t distribution with $n - 1 = 26$ degrees of freedom. Note that we use t rather than z , since we don't know the value of the population standard deviation σ . We only know the sample standard deviation s .

- (b) If we took repeated samples of 27 trees and calculated the interval in a similar manner, 98% of such intervals would contain the true mean height of all trees in the forest.
- (c) Let $\alpha = 0.02$.

We are testing the hypotheses $H_0: \mu = 30$ vs. $H_a: \mu < 30$.

We will reject H_0 if the P-value $\leq \alpha = 0.02$.

The test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{28.4 - 30}{5.3/\sqrt{27}} = -1.57$.

The P-value is $P(T(27) \leq -1.57) = P(T(27) \geq 1.57)$ by the symmetry of the t distributions. From Table 2 we see

$$P(T(26) \geq 1.315) = 0.10 \quad \text{and} \quad P(T(26) \geq 1.706) = 0.05$$

Since $1.315 < 1.57 < 1.706$, the P-value is between 0.05 and 0.10.

Since the P-value $> \alpha = 0.02$, we fail to reject H_0 . At the 2% level of significance, we have insufficient evidence that the true mean height of trees in the forest is less than 30 feet.

- (d) If the true mean height of all trees in the forest was 30 feet, the probability of observing a value of $\bar{x}$ at least as low as 28.4 feet would be between 0.05 and 0.10.
- (e) Confidence intervals can only be used to conduct two-sided hypothesis tests. Since this was a one-sided test, the confidence interval could **not** have been used.

Sample Final Exam 2 – Part A

Question 1

The median is in position $(n + 1)/2 = (25 + 1)/2 = 13$, so the median is 30.

The first quartile is the median of all data values lower in position than the median, so $Q1$ is in position $(12 + 1)/2 = 6.5$, i.e., the average of the 6th and 7th values. Therefore, $Q1 = (18 + 19)/2 = 18.5$.

The third quartile is the median of all data values higher in position than the median, so $Q3$ is in position 6.5 above the median (or equivalently, 6.5 positions down from the maximum). Therefore, $Q3 = (48 + 50)/2 = 49$.

The interquartile range is $Q3 - Q1 = 49 - 18.5 = 30.5$.

Question 2

The residual for Matt Nichols is

$$y_5 - \hat{y}_5 = 1757 - (-53.51 + 12.53(149)) = 1757 - 1813.46 = -56.46$$

Question 3

$$\begin{aligned} P(15 < X < 25) &= P\left(\frac{15 - 18}{3} < Z < \frac{25 - 18}{3}\right) = P(-1.00 < Z < 2.33) \\ &= P(Z < 2.33) - P(Z < -1.00) = 0.9901 - 0.1587 = 0.8314 \end{aligned}$$

Question 4

We know $P(X < 320) = 0.1000$. We find the value z such that $P(Z < z) = 0.1000$ to be $z = -1.28$. Now we use the z -score formula to solve for σ :

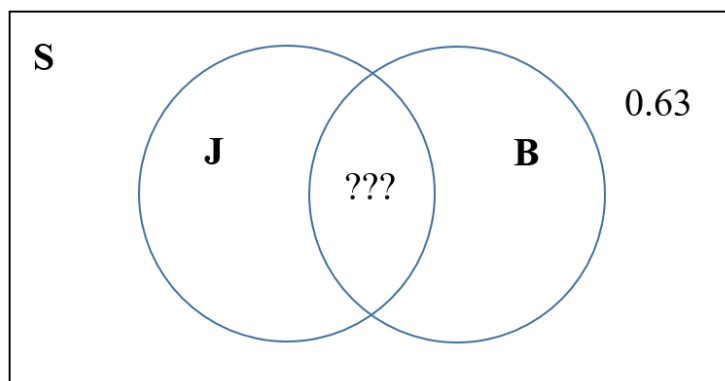
$$z = \frac{x - \mu}{\sigma} \Rightarrow \sigma = \frac{x - \mu}{z} = \frac{320 - 384}{-1.28} = 50$$

Question 5

$$P(\text{Pass}) = 1 - P(\text{Fail}) = 1 - P(D \text{ or } F) = 1 - (0.09 + 0.07) = 1 - 0.16 = 0.84$$

Question 6

We know the probability that someone has been to see neither team is 0.63. We see from the Venn diagram below that it follows that $P(J \cup B) = 1 - 0.63 = 0.37$.



We are looking for $P(J \cap B)$. We know that

$$P(J \cup B) = P(J) + P(B) - P(J \cap B)$$

$$\Rightarrow P(J \cap B) = P(J) + P(B) - P(J \cup B) = 0.32 + 0.25 - 0.37 = 0.20$$

Question 7

The sample space is $S = \{BB, BG, GB, GG\}$. (Order matters!)

The event that exactly one child is a boy is $A = \{BG, GB\}$. The complement of this event is all outcomes in the same space not contained in A , i.e., $A^c = \{BB, GG\}$. This is the event that the couple has either zero or two girls.

Question 8

We have a fixed number of trials (we are selecting $n = 4$ cards). There are only two outcomes of interest for each selection – either the card is blue or it is not. Since cards are selected with replacement, trials are independent, and there is a constant probability of success ($p = 3/10 = 0.3$). Therefore, X has a binomial distribution with parameters $n = 4$ and $p = 0.3$.

Question 9

Two events are mutually exclusive if they CANNOT both happen at the same time (i.e., there are no outcomes in the sample space satisfying the definition of both events). If we can find any outcome satisfying both events, then they are not mutually exclusive. We define the event A = first card is the only red card.

If A happens, is it possible that the second selected card is blue? YES! One outcome that satisfies both A and this event is $RBBG$. Therefore, A and this event are **not** mutually exclusive.

If A happens, is it possible that no green cards are selected? YES! There is one outcome that satisfies both A and this event, i.e., $RBBB$. Therefore, A and this event are **not** mutually exclusive.

If A happens, is it possible that the third selected card is the only green card? YES! There is one outcome that satisfies both A and this event, i.e., $RBGB$. Therefore, A and this event are **not** mutually exclusive.

If A happens, is it possible that the same number of red and blue cards are selected? YES! One outcome that satisfies both A and this event is $RBGG$ – there is one red and one blue.

If A happens, is it possible that the same number of blue and green cards are selected? NO! We know the first card is the ONLY red card selected. This means the last three cards are blue and/or green. There is no way to have an equal number of blue and green cards (we can't have 1.5 of each). Therefore, this event and A **are** mutually exclusive.

Question 10

$$P(\text{first green on fifth draw}) = P(G^c G^c G^c G^c G) = (0.8)(0.8)(0.8)(0.8)(0.2) = 0.0819.$$

The probability any selected card is green is $2/10 = 0.2$. We can multiply probabilities because cards are selected with replacement, so trials are independent.

Question 11

Let X be the number of flares that light. There is a fixed number of trials ($n = 5$), and there are only two outcomes of interest for each trial – the flare either lights or it doesn't. Trials are independent with constant probability of success ($p = 0.93$). Therefore,

$$P(X = 4) = \binom{5}{4} (0.93)^4 (0.07)^1 = 0.2618$$

Question 12

Let X be the number of the 500 students who live in dorms. Then X has a binomial distribution with parameters $n = 500$ and $p = 0.11$. Since the sample size is very large, we can use the normal approximation (you can check that np and $n(1 - p)$ are both greater than 10):

$$\begin{aligned} P(\hat{p} \geq 0.14) &\approx P\left(Z \geq \frac{0.14 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(Z \geq \frac{0.14 - 0.11}{\sqrt{\frac{0.11(0.89)}{500}}}\right) \\ &= P(Z \geq 2.14) = 1 - P(Z < 2.14) = 1 - 0.9838 = 0.0162 \end{aligned}$$

Question 13

Note that we are taking a large sample from a population and plotting the INDIVIDUAL VALUES. This question has nothing to do with a normal distribution or the Central Limit Theorem. (We are NOT plotting sample means.) All we are doing is taking a large sample from the population and plotting the values, so we expect to get something similar to the original population distribution. Think about it the other way around – if we had no idea what the population looked like, and we took a sample of 2500 values and plotted them....then if we saw a distribution that was skewed to the left, it would be logical to conclude the population these values came from is likely skewed to the left.

Question 14

The part of this question that confuses some people is that we are given the variance σ^2 , and not the standard deviation σ . (Note that when we say the variance is 36 months squared, the “squared” applies only to the units, not to the 36.) The variance is $\sigma^2 = 36$, so the standard deviation is $\sigma = \sqrt{36} = 6$. Therefore,

$$P(\bar{X} < 52) = P\left(Z < \frac{52 - 54}{6/\sqrt{50}}\right) = P(Z < -2.36) = 0.0091$$

Question 15

$$\begin{aligned} P(99 < \bar{X} < 101) &= P\left(\frac{99 - 100}{8/\sqrt{64}} < Z < \frac{101 - 100}{8/\sqrt{64}}\right) = P(-1.00 < Z < 1.00) \\ &= P(Z < 1.00) - P(Z < -1.00) = 0.8413 - 0.1587 = 0.6826 \end{aligned}$$

Question 16

This question isn't really about confidence intervals – it's a binomial question. Let X be the number of statisticians who get an interval containing μ . There is a fixed number of trials ($n = 10$), and there are only two outcomes of interest for each trial (either the interval contains μ or it doesn't). Trials are independent with constant probability of success (the probability of any given interval containing μ is $p = 0.9$). Therefore, X has a binomial distribution with parameters $n = 10$ and $p = 0.9$, and so

$$P(X = 8) = \binom{10}{8} (0.9)^8 (0.1)^2 = 0.1937$$

Question 17

The question says we want to estimate μ to within 4 mm Hg. This tells us that the margin of error $m = 4$. We are also given $n = 25$ and $\sigma = 8.6$. We know the margin of error is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

In order to determine the confidence level, we must solve for z^* :

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad z^* = \frac{m\sqrt{n}}{\sigma} = \frac{4\sqrt{25}}{8.6} = 2.326$$

From the bottom of Table 2, we see this critical value z^* corresponds to a 98% confidence interval.

Question 18

We need to find the critical value z^* for an 85% confidence interval. Draw the standard normal curve for Z , centred at $\mu = 0$. Now draw lines at the values $-z^*$ and z^* on either side of the mean (equally far in either direction), such that the area between $-z^*$ and z^* is 0.85. In order to find the value z^* , we need to find the total area to the left of z^* (so we can use the table to look it up). From the picture, you see that this is equal to the area to the left of $-z^*$ plus the area between $-z^*$ and z^* . We already know $P(-z^* < Z < z^*) = 0.8500$, so we need to find $P(Z < -z^*)$. We know the central area (between $-z^*$ and z^*) is 0.8500, so the total area of the two tails (i.e., below $-z^*$ and above z^*) is $1 - 0.8500 = 0.1500$. That means each tail has an area equal to half this value, $0.1500/2 = 0.0750$. Therefore, the total area to the left of z^* is

$$P(Z < z^*) = P(Z < -z^*) + P(-z^* < Z < z^*) = 0.0750 + 0.8500 = 0.9250.$$

Using the normal table (going backwards), we find that the value of z^* with 0.9250 area to its left is $z^* = 1.44$. (We take the closest value in the table.)

Therefore, an 85% confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 3700 \pm 1.44 \left(\frac{500}{\sqrt{25}} \right) = 3700 \pm 144 = (3556, 3844)$$

Question 19

We are testing $H_0: \mu = 900$ vs. $H_a: \mu > 900$.

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{920 - 900}{80/\sqrt{64}} = 2.00$.

The P-value is $P(Z \geq 2.00) = 1 - P(Z < 2.00) = 1 - 0.9772 = 0.0228$.

Since the P-value $< \alpha = 0.03$, we reject H_0 . At the 3% level of significance, we have sufficient evidence to conclude that the true mean lifetime of the new bulbs is greater than 900 hours. Therefore, the university will purchase the new bulbs.

Question 20

In order to use a confidence interval to conduct a test, the test must be two-sided (which this one is) AND the confidence level and α must add up to 100%.

We are testing $H_0: \mu = 7.0$ vs. $H_a: \mu \neq 7.0$. We have a 98% confidence interval for μ , so if we want to use the interval to conduct the test, we must use a 2% level of significance (since a hypothesis test with $\alpha = 0.02$ and a 98% confidence interval both leave out 1% area in each tail of the distribution). This eliminates options (B), (C) and (E).

When using a confidence interval to conduct a test, we must check if μ_0 (i.e., the value of μ under the null hypothesis) is contained in the interval. Here we are testing whether the mean differs from 7.0, so $\mu_0 = 7.0$. Since the value 7.0 is contained in the 98% confidence interval, we fail to reject H_0 at the 2% level of significance.

Question 21

The formula for the confidence interval is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Note that we must use t rather than z , as the standard deviation given in the question is the sample standard deviation s . In order to determine the confidence level, we need to find the critical value t^* . We know that $\bar{x}$ is at the centre of the confidence interval, so $\bar{x}$ is the average of the lower and upper confidence limits, i.e.,

$$\bar{x} = \frac{126.41 + 153.59}{2} = 140$$

We also know that $s = 19$ and $n = 10$. We can find t^* using either the lower or upper limit of the confidence interval. We will use the upper limit:

$$\bar{x} + t^* \frac{\sigma}{\sqrt{n}} = 153.59$$

$$\Rightarrow 140 + t^* \left(\frac{19}{\sqrt{10}} \right) = 153.59$$

$$\Rightarrow z^* = (153.59 - 140) \left(\frac{\sqrt{10}}{19} \right) = 2.262$$

We now go along the row for $n - 1 = 9$ degrees of freedom (don't use the last row of the table – this is for z^* , not t^*). From the row for 9 d.f., we see that $t^* = 2.262$ corresponds to a 95% confidence interval.

Question 22

We are testing $H_0: \mu = 4.8$ vs. $H_a: \mu > 4.8$. Note that this is an upper-tailed test because we are testing whether the home is **unsafe**. We are told that the home is safe if $\mu \leq 4.8$, so it is unsafe if $\mu > 4.8$. Note that in a hypothesis test, we are always looking for evidence to support H_a .

The test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.0 - 4.8}{0.5/\sqrt{8}} = 1.13$.

Note that this is a t test rather than a z test, since we were only given the sample standard deviation s . (It says “these readings...have a standard deviation of 0.5”, referring to the sample.)

The P-value is $P(T(7) \geq 1.13)$. From Table 2, we see that

$$P(T(7) \geq 1.119) = 0.15 \quad \text{and} \quad P(T(7) \geq 1.415) = 0.10$$

Since $1.119 < t = 1.13 < 1.415$, the P-value is between 0.10 and 0.15.

Question 23

Since the P-value $> \alpha = 0.01$, we fail to reject H_0 . We have insufficient evidence that the home is unsafe, i.e., that H_a is true. (Note that we do not conclude the home is safe. There is just not enough evidence that it is unsafe.)

Question 24

The heights of **all** Jets players are given. There is no need for a hypothesis test. In a hypothesis test, we use information from a sample to draw conclusions about a larger population. In this case, we already have the whole population. If you calculated the average height of these 24 players, this is already the value of μ . There is no need to estimate it or conduct a hypothesis test.

Question 25

This is a matched-pairs study. For any University A student, high school GPA and university GPA are **DEPENDENT** (i.e., related), so Statement (I) is false. For any two students, high school GPAs are independent (unrelated), so Statement (II) is true. In the matched pairs t procedures, it is not enough to assume each variable separately has a normal distribution. We must assume the **DIFFERENCES** follow a normal distribution, so Statement (III) is true and Statement (IV) is false.

Question 26

The information about high school GPAs and university GPAs separately are “extra information” – we only care about the differences. A 95% confidence interval for μ_d is

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}} = 0.17 \pm 2.447 \left(\frac{0.47}{\sqrt{7}} \right) = 0.17 \pm 0.43 = (-0.26, 0.60)$$

where $t^* = 2.447$ is the upper 0.025 critical value for the t distribution with $n - 1 = 6$ degrees of freedom.

Question 27

This is a matched pairs t test. We are testing $H_0: \mu_d = 0$ vs. $H_a: \mu_d \neq 0$.

The test statistic is $t = \frac{\bar{x}_d}{s_d/\sqrt{n}} = \frac{0.26}{0.33/\sqrt{8}} = 2.23$.

The P-value is $2P(T(7) \geq 2.23)$. We see from Table 2 that

$$P(T(7) \geq 1.895) = 0.05 \quad \text{and} \quad P(T(7) \geq 2.365) = 0.025$$

Since $1.895 < t = 2.23 < 2.365$, it follows that $P(T(7) \geq 2.23)$ is between 0.025 and 0.05. But the P-value is double this probability, so the P-value is between 0.05 and 0.10.

Question 28

The required sample size is $n = \left(\frac{z^*}{m} \right)^2 p^*(1 - p^*) = \left(\frac{1.960}{0.06} \right)^2 (0.5)(0.5) = 266.78 \approx 267$.

In sample size calculations for proportions, unless the question specifically states some other value of p , we always use $p^* = 0.5$ in our calculation. This gives us the maximum sample size we could possibly need to achieve our margin of error, regardless of the true value of p .

Question 29

A 90% confidence interval for p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.837 \pm 1.645 \sqrt{\frac{0.837(0.163)}{1000}} = 0.837 \pm 0.019 = (0.818, 0.856)$$

Question 30

We are testing $H_0: p = 0.30$ vs. $H_a: p > 0.30$. We calculate $\hat{p} = \frac{70}{175} = 0.40$.

The test statistic is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.40 - 0.30}{\sqrt{\frac{0.30(0.70)}{175}}}.$

Sample Final Exam 2 – Part B

1. (a)
 - i. The factors in an experiment are the explanatory variables. There are two factors in this experiment – brand of toothpaste and type of toothbrush.
 - ii. Brand of toothpaste has two factor levels – Crest and Colgate.
Type of toothbrush has two factor levels – regular and electric.
 - iii. The treatments are the combinations of factor levels. There are $2 \times 2 = 4$ treatments – Crest/regular, Crest/electric, Colgate/regular, Colgate/electric.
 - iv. The response variable is the health of the patients' teeth.
 - v. The blocking variable is age. (The blocks are the children and the adults.)
 - (b) Control is achieved through the comparison of the various treatments. (Note that we do NOT need a control group. Note also that blocking is also another form of control. In this case, by using a randomized block design, we eliminate the possibility of age becoming a lurking variable.)
 - (c) Since this is a properly conducted experiment using the principles of randomization, replication and control, we **can** say that the treatment was likely the cause.
2. (a) The sample space of outcomes is shown below, as well as the probability of each outcome:

Outcome	Probability
WWW	$(0.6)(0.7)(0.8) = 0.336$
WWB	$(0.6)(0.7)(0.2) = 0.084$
WMW	$(0.6)(0.3)(0.8) = 0.144$
WMB	$(0.6)(0.3)(0.2) = 0.036$
PWW	$(0.4)(0.7)(0.8) = 0.224$
PWB	$(0.4)(0.7)(0.2) = 0.056$
PMW	$(0.4)(0.3)(0.8) = 0.096$
PMB	$(0.4)(0.3)(0.2) = 0.024$

Note that we multiply probabilities, since the outcome of one game is independent of any other.

- (b) Let X be the number of games won by the Jets. The variable X can take the values 0, 1, 2 or 3. We find the probability distribution of X :

$$P(X = 0) = P(PMB) = 0.024$$

$$P(X = 1) = P(WMB) + P(PWB) + P(PMW) = 0.036 + 0.056 + 0.096 = 0.188$$

$$P(X = 2) = P(WWB) + P(WMW) + P(PWW) = 0.084 + 0.144 + 0.224 = 0.452$$

$$P(X = 3) = P(WWW) = 0.336$$

The probability distribution of X is summarized in the table below:

x	0	1	2	3
$P(X = x)$	0.024	0.188	0.452	0.336

3. (a) We want to find the value x such that $P(X > x) = 0.1200$. We first find the value z such that $P(Z > z) = 0.1200 \Rightarrow P(Z < z) = 0.8800$. This corresponds to the value $z = 1.175$. (Note that we usually choose the closest entry in the table. However, in this case the value 0.8800 falls exactly half way between 0.8790 and 0.8810, so we take the average of $z = 1.17$ and $z = 1.18$.)

We now use the z -score formula to solve for x :

$$z = \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma = 5.0 + 1.175(0.2) = 5.235$$

- (b) Since the population distribution is normal, we can calculate this probability, even though the sample size is low. (If X is normal, the distribution of $\bar{X}$ is exactly normal for any sample size n .)

$$\begin{aligned} P(\bar{X} > 5.1) &= P\left(Z > \frac{5.1 - 5.0}{0.2/\sqrt{4}}\right) = P(Z > 1.00) = 1 - P(Z \leq 1.00) \\ &= 1 - 0.8413 = 0.1587 \end{aligned}$$

- (c) $P(\text{Total} < 248) = P\left(\bar{X} < \frac{248}{50}\right) = P(\bar{X} < 4.96) = P\left(Z < \frac{4.96 - 5.0}{0.2/\sqrt{50}}\right)$
 $= P(Z < -1.41) = 0.0793$

- (d) The probability is exact. Since the population distribution is normal, the sampling distribution of $\bar{X}$ is exactly normal.

4. (a) A 95% confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 3.12 \pm 1.96 \left(\frac{0.47}{\sqrt{20}} \right) = 3.12 \pm 0.21 = (2.91, 3.33)$$

- (b) If we took repeated samples of 20 students and calculated the interval in a similar manner, 95% of such intervals would contain the true mean GPA of all students at the university.
- (c) Let $\alpha = 0.05$.

We are testing the hypotheses $H_0: \mu = 3.00$ vs. $H_a: \mu \neq 3.00$.

We will reject H_0 if the P-value $\leq \alpha = 0.05$.

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.12 - 3.00}{0.47/\sqrt{20}} = 1.14$.

The P-value is $2P(Z \geq 1.14) = 2(1 - 0.8729) = 2(0.1271) = 0.2542$.

Since the P-value $> \alpha = 0.05$, we fail to reject H_0 . At the 5% level of significance, we have insufficient evidence that the true mean GPA differs from 3.00.

- (d) If the true mean GPA was 3.00, the probability of observing a value of $\bar{x}$ at least as extreme as 3.12 would be 0.2542.
- (e) Since this is a two-sided test, and since the level of significance and the confidence level add up to 100%, the confidence interval can be used to conduct the test. Since $\mu_0 = 3.00$ is contained in the 95% confidence interval for μ , we fail to reject H_0 at the 5% level of significance.

Sample Final Exam 3 – Part A

Question 1

We calculate the fences:

$$LF = Q1 - 1.5(IQR) = 174 - 1.5(181 - 174) = 174 - 10.5 = 163.5$$

$$UF = Q3 + 1.5(IQR) = 181 + 1.5(181 - 174) = 181 + 10.5 = 191.5$$

Any data value less than 163.5 or greater than 181.5 is an outlier. Only Kyle's height is an outlier.

Question 2

We are told that 58% of the variation in temperature can be explained by its regression on the height of a mountain. This is, by definition, r^2 . We know $r^2 = 0.58$, and so the correlation is $r = -\sqrt{0.58} = -0.76$. Note that we take the negative square root of r^2 , as we know the relationship between the height of a mountain and temperature is negative.

Question 3

The factors in an experiment are the explanatory variables. There are two factors in this experiment – temperature and wind speed.

Temperature has two factor levels: -20°C and -30°C . Wind speed has two factor levels: 30 km/h and 40 km/h.

Question 4

The sample space is $S = \{G, SG, CG, SCG, CSG\}$. There are 5 outcomes.

Question 5

We know that $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B) = 0.7354$.

Note that $P(A \cap B) = P(A)P(B)$ because we know A and B are independent. Now we plug in $P(A) = 0.37$ and solve for $P(B)$:

$$0.37 + P(B) - 0.37P(B) = 0.7354 \quad \Rightarrow \quad 0.37 + 0.63P(B) = 0.7354$$

$$0.63P(B) = 0.7354 - 0.37 = 0.3654 \quad \Rightarrow \quad P(B) = \frac{0.3654}{0.63} = 0.58$$

Question 6

Let's call the probability of getting any odd number k . Then the probability of any even number is $2k$. The probability distribution is shown below:

Number	1	2	3	4	5	6
Probability	k	$2k$	k	$2k$	k	$2k$

We know all probabilities must add up to 1, so we solve for k :

$$k + 2k + k + 2k + k + 2k = 1 \quad \Rightarrow \quad 9k = 1 \quad \Rightarrow \quad k = 1/9$$

The probability of rolling a number greater than 3 is therefore

$$P(4) + P(5) + P(6) = 2k + k + 2k = 5k = 5(1/9) = 5/9$$

Question 7

The probability Joe wins is

$$\begin{aligned} &P(\text{Joe picks Rock and Tom picks Scissors}) \\ &+ P(\text{Joe picks Paper and Tom picks Rock}) \\ &+ P(\text{Joe picks Scissors and Tom picks Paper}) \\ &= (0.4)(0.3) + (0.2)(0.5) + (0.4)(0.2) = 0.30 \end{aligned}$$

Question 8

Statement (A) is true. If A and B were mutually exclusive, then we would have $P(A \cap B) = 0$, and therefore $P(A \cup B) = P(A) + P(B)$. But that is impossible in this case, because $P(A \cup B)$ would be 1.1, and a probability can't be greater than 1. Another way to see this is to draw a Venn diagram. Draw a circle for $P(A)$ that takes up 60% of the rectangle. In order for A and B to be mutually exclusive, we would have to be able to draw a circle for $P(B)$ that doesn't overlap the A circle. But since the B circle takes up 50% of the rectangle, it's not possible for it not to overlap with the A circle, which takes up 60% of the area. Therefore, A and B cannot be mutually exclusive.

Statement (B) is true. If A and B are independent, then

$$P(A \cap B) = P(A)P(B) = (0.4)(0.6) = 0.24$$

Statement (C) is true. We know

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.6 - P(A \cap B) = 0.72$$

So we know $P(A \cap B) = 0.18$. But this is equal to $P(A)P(B) = 0.3(0.6) = 0.18$, so A and B are independent.

Statement (D) is false. It is easy to illustrate a counter-example. Draw a Venn diagram where the B and C circles overlap, but the A circle doesn't overlap either one of them. Therefore, A and B are mutually exclusive and A and C are mutually exclusive, but B and C are **not** mutually exclusive.

Statement (E) is true. If two events are mutually exclusive, $P(A \cap B) = 0$. If two events are independent, $P(A \cap B) = P(A)P(B)$, which is obviously positive (as long as $P(A)$ and $P(B)$ are both greater than zero). Another way to think of it...if two events are independent, then whether A happens tells us nothing about whether B will happen. If A and B are mutually exclusive, knowing A happens obviously tells us something about whether B will happen – it can't happen! Therefore the two events can't be independent if they are mutually exclusive.

Question 9

A binomial experiment is one for which all four of the following four conditions are satisfied:

1. There is a fixed number of trials n .
2. There are only two possible outcomes of interest for each trial.
3. Trials are independent.
4. The probability of success p is constant for each trial.

The variable in (A) is **not** binomial. We are simply counting the total number of dots; we are not counting the number of successes in trials with only two outcomes of interest.

The variable in (B) is **not** binomial. There is no constant probability of success p . The probability that it snows on a Saturday in January is obviously greater than the probability it snows on a Saturday in July.

The variable in (C) is **not** binomial. Outcomes are not independent. If we know one house has not gotten their newspaper on time, we know the paper boy hasn't come by yet, and so their neighbour probably hasn't gotten their paper either.

The variable in (D) is **not** binomial. There is no fixed number of trials.

The variable in (E) **is** binomial. There are a fixed number of trials ($n = 30$) and there are only two outcomes of interest for each question (even though there are five choices for each question, we are only interested in whether the student gets the question right or wrong). The student randomly guesses the answer for each question, so outcomes are independent with constant probability of success ($p = 1/5 = 0.2$ since there are five choices A, B, C, D and E for each question).

Question 10

Let X be the number of times you win during the week. There is a fixed number of trials ($n = 7$) and there are only two outcomes of interest for each trial (either you win or you lose). Whether you win one day is independent of any other day and there is a constant probability of success ($p = 0.167$). The probability you win at least twice is

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) = 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left[\binom{7}{0} (0.167)^0 (0.833)^7 + \binom{7}{1} (0.167)^1 (0.833)^6 \right] \\ &= 1 - (0.2783 + 0.3906) = 1 - 0.6689 = 0.3311 \end{aligned}$$

Question 11

$$P(\bar{X} > 830) = P\left(Z > \frac{830 - 828}{4/\sqrt{10}}\right) = P(Z > 1.58)$$

$$= 1 - P(Z \leq 1.58) = 1 - 0.9429 = 0.0571$$

Question 12

According to the 68-95-99.7 rule, there is an approximate 99.7% chance that a normally distributed variable falls within 3 standard deviations of its mean. The mean of $\bar{X}$ is μ and the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$, so there is an approximate 99.7% chance that $\bar{X}$ falls between

$$\mu - 3\frac{\sigma}{\sqrt{n}} = 828 - 3\left(\frac{4}{\sqrt{16}}\right) = 828 - 3 = 825$$

and

$$\mu + 3\frac{\sigma}{\sqrt{n}} = 828 + 3\left(\frac{4}{\sqrt{16}}\right) = 828 + 3 = 831$$

Question 13

We want to find the value x such that $P(X < x) = 0.0400$. First we find the value z such that $P(Z < z) = 0.0400$ to be $z = -1.75$. Now we use the z -score formula to solve for x :

$$z = \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma = 828 + (-1.75)(4) = 828 - 7 = 821$$

Question 14

Let X be the number of the 300 arrows that hit the bullseye. Then X has a binomial distribution with parameters $n = 300$ and $p = 0.78$. Since the sample size is very large, we can use the normal approximation (you can check that np and $n(1 - p)$ are both greater than 10):

$$P\left(\hat{p} \geq \frac{240}{300}\right) = P(\hat{p} \geq 0.80) \approx P\left(Z \geq \frac{0.80 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(Z \geq \frac{0.80 - 0.78}{\sqrt{\frac{0.78(0.22)}{300}}}\right)$$

$$= P(Z \geq 0.84) = 1 - P(Z < 0.84) = 1 - 0.7995 = 0.2005$$

Question 15

Answer (E) is correct.

Question 16

$P(\bar{X} = 339.5) = 0$, since $\bar{X}$ is a continuous variable. This corresponds to the area of a line under the density curve, which is zero.

Question 17

We convert 4.1 litres to 4100 millilitres and we calculate

$$P(\text{Total} > 4100) = P\left(\bar{X} > \frac{4100}{12}\right) = P(\bar{X} > 341.667) = P\left(Z > \frac{341.667 - 341}{3/\sqrt{12}}\right)$$

$$= P(Z > 0.77) = 1 - P(Z \leq 0.77) = 1 - 0.7794 = 0.2206$$

Question 18

We want to find the value $\bar{x}$ such that $P(\bar{X} < \bar{x}) = 0.0250$. We first find the value z such that $P(Z < z) = 0.0250$ to be $z = -1.96$. Now we use the z formula to solve for $\bar{x}$ as follows:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \bar{x} = \mu + z \frac{\sigma}{\sqrt{n}} = 450 + (-1.96) \left(\frac{12}{\sqrt{5}}\right) = 439.48$$

Question 19

Since the population distribution is normal, the sampling distribution of $\bar{X}$ is exactly normal with mean $\mu = 846$ and standard deviation $\sigma/\sqrt{n} = 160/\sqrt{4} = 80$.

Question 20

The populations are both large, so the actual size of the population is irrelevant. It is not the fraction of students sampled that matters (as long as the fraction is small) – it is the actual sample size.

The standard deviation of $\bar{X}$ for University A is $\sigma/\sqrt{500}$ and the standard deviation of $\bar{X}$ for University B is $\sigma/\sqrt{600}$. Therefore, since the standard deviation for University B will be lower than that for University A, and since the confidence interval is $\bar{x} \pm 1.96(\text{std. dev.})$, the interval for University B will be narrower. As such, Statement (I) is false and Statement (II) is true.

The higher the confidence level, the wider the interval, so a 90% confidence interval would be **narrower** than a 95% confidence interval. Statement (III) is false.

In order to cut the length of the interval in half, we would need $2^2 = 4$ times the sample size. Doubling the sample size does **not** cut the margin of error in half. In general, if we want to decrease the margin of error by a factor of k , we need k^2 times the original sample size. Statement (IV) is false.

Question 21

For a fixed confidence level, when we **decrease** the margin of error by a factor of k , the required sample size is k^2 times **greater**. Therefore, when we **increase** the margin of error by a factor of k , the required sample size is k^2 times **smaller**. Here, $k = 2$, i.e., we are doubling the margin of error (increasing it by a factor of 2). Therefore, we need a sample size $2^2 = 4$ times smaller, i.e., $n = 120/4 = 30$. This can be seen by looking at the sample size formula:

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

When the margin of error m becomes 2 times larger, the fraction becomes 2 times **smaller**. Then we square the quantity inside the parentheses, so the required sample size is $2^2 = 4$ times smaller.

Question 22

It is the sample size that affects the margin of error (and hence the length of the confidence interval), and not the population size (as long as the populations are large). If we want to make a confidence interval for Montreal with the same confidence level and the same margin of error as the interval for Hamilton, we need the same sample size.

Question 23

We are conducting a two-sided test at $\alpha = 0.05$. We reject H_0 , so we know the P-value $< \alpha = 0.05$.

Consider Statement (I). We reject H_0 if the P-value $< \alpha$. We know the P-value is less than 0.05, since we rejected H_0 at $\alpha = 0.05$. The P-value is obviously less than 0.10 then. Therefore, we would reject H_0 at $\alpha = 0.10$. Statement (I) is true.

Note that if we were given the information the other way around – i.e., if we rejected H_0 at a 10% level of significance, we would know the P-value is less than 0.10, so would we reject H_0 at $\alpha = 0.05$? We don't know! The P-value is less than 0.10 but we don't know if it's less than 0.05.

Consider Statement (II). We know the P-value for a two-sided test is DOUBLE the P-value for a one-sided test. The P-value of the original two-sided test is two times the area to the right of $\bar{x} = 33$. The P-value in (II) would only be half of this value. Since we know the two-sided P-value is less than 0.05, the one-sided P-value must also be less than 0.05. Statement (II) must be true.

Note that if we were given the information the other way around – i.e., if we rejected H_0 for the one-sided test and we wanted to know whether we would reject H_0 for the two-sided test – the answer would be WE DON'T KNOW. If we rejected the one-sided test, we would know the P-value is less than 0.05. The P-value of the two-sided test is double this original P-value. Would it still be less than 0.05? There's no way to know!

Consider Statement (III). A sample mean of $\bar{x} = 33$ was far enough out in the tail to reject H_0 and conclude that $\mu \neq 30$. So is $\bar{x} = 33$ far enough in the tail to conclude that $\mu \neq 31$? We don't know! Statement (III) is not necessarily true.

Suppose the question was asked the other way around. If we had enough evidence to conclude that $\mu \neq 31$ based on a sample mean of $\bar{x} = 33$, then it is definitely extreme enough to conclude that $\mu \neq 30$.

Question 24

We are given both the values of the population standard deviation σ and the sample standard deviation s . When deciding whether to use z or t , we ask, “Do I know σ ?”. If we do, we use z . If we don’t, we have to estimate σ with s and use t .

In this example, the sample standard deviation is just extra information. We know σ , so we use z . If we know the true population standard deviation, there’s no need to estimate it.

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{598.8 - 600}{2/\sqrt{25}} = -3.00$.

Question 25

We are testing the hypotheses $H_0: \mu = 350$ vs. $H_a: \mu \neq 350$.

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{348.7 - 350}{2.8/\sqrt{16}} = -1.86$.

The P-value is $2P(Z \leq -1.86) = 2(0.0314) = 0.0628$.

Since the P-value $> \alpha = 0.05$, we fail to reject H_0 .

Question 26

In this question, we consider using a confidence interval to conduct a hypothesis test. In order to use a confidence interval to conduct a test, it must be a two-sided test, and the confidence level and the level of significance must add up to 100%. If the value μ_0 falls in the corresponding interval, we fail to reject H_0 . If μ_0 falls outside the interval, we reject H_0 .

If we wanted to use a 90% confidence interval to conduct the test, we would have to use a 10% level of significance. Since $\mu_0 = 50$ falls outside the 90% confidence interval, we know we would reject H_0 at $\alpha = 0.10$, i.e., the P-value must be less than 0.10.

If we wanted to use a 96% confidence interval to conduct the test, we would have to use a 4% level of significance. Since $\mu_0 = 50$ falls inside the 96% confidence interval, we know we would fail to reject H_0 at $\alpha = 0.04$, i.e., the P-value must be greater than 0.04.

Therefore, we know the P-value must be between 0.04 and 0.10.

Question 27

For any individual, the scores for Coke and Pepsi are dependent (related), since they come from the same individual. Statement (I) is false and Statement (II) is true.

In a matched pairs t test, it is not enough to assume each variable separately follows a normal distribution. We must assume that the **differences** follow a normal distribution. Statement (III) is false and Statement (IV) is true.

Question 28

In a matched pairs t test, we are testing the null hypothesis $H_0: \mu_d = 0$. The question simply asks if people prefer one cola over the other, and does not mention a specific direction. As such, the alternative hypothesis is two-sided, $H_a: \mu_d \neq 0$.

Question 29

The test statistic is $t = \frac{\bar{x}_d}{s_d/\sqrt{n}} = \frac{-5}{13/\sqrt{20}} = -1.72$.

Question 30

We always calculate the P-value assuming H_0 is true. Statement (A) is true.

Statement (B) is true. This is why we reject H_0 for LOW P-values.

Statement (C) is true. We always reject H_0 if the P-value $\leq \alpha$.

Statement (D) is **false**. A hypothesis is designed to assess the strength of evidence **against** H_0 (and in favour of H_a). This is why we either reject H_0 , or we fail to reject H_0 (but never “accept H_0 ”).

Statement (E) is true. In a hypothesis test, we are using information from a sample to draw conclusions about a whole population. The only way this can be accurate is if the sample is representative of the population. We try to ensure we get representative samples by selecting them randomly.

Sample Final Exam 3 – Part B

1. (a) The fraction of the variation in Y explained by its regression on X is, by definition, equal to r^2 . Here, $r^2 = (-0.869)^2 = 0.755$.

(b) The slope is $b_1 = r \frac{s_y}{s_x} = (-0.869) \left(\frac{25.07}{2.61} \right) = -8.347$.

The intercept is $b_0 = \bar{y} - b_1 \bar{x} = 64 - (-8.347)(4) = 97.388$.

The equation of the least squares regression line is $\hat{y} = 97.388 - 8.347x$.

- (c) The predicted score for a subject who drinks 5 drinks is

$$\hat{y} = 97.388 - 8.347(5) = 55.653$$

The predicted score for a subject who drinks 11 drinks is

$$\hat{y} = 97.388 - 8.347(11) = 5.571$$

- (d) The predicted value for the subject who drank 5 drinks is more reliable. This is because $x = 5$ is within our range of x -values in the sample, while $x = 11$ falls outside our range of data. Predicting the score for the subject who drank 11 drinks is extrapolation.

2. (a) We know $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Therefore,

$$P(B) = P(A \cup B) - P(A) + P(A \cap B) = 0.61 - 0.43 + 0.19 = 0.37$$

- (b) We know A and B are not mutually exclusive, since $P(A \cap B) > 0$. We know B and C are not mutually exclusive, since $P(B \cap C) > 0$. To determine whether A and C are mutually exclusive, we have to find $P(A \cap C)$. We know that $P(A \cup C) = P(A) + P(C) - P(A \cap C)$. Therefore,

$$P(A \cap C) = P(A) + P(C) - P(A \cup C) = 0.71 - 0.43 - 0.28 = 0$$

Since $P(A \cap C) = 0$, it follows that A and C **are** mutually exclusive.

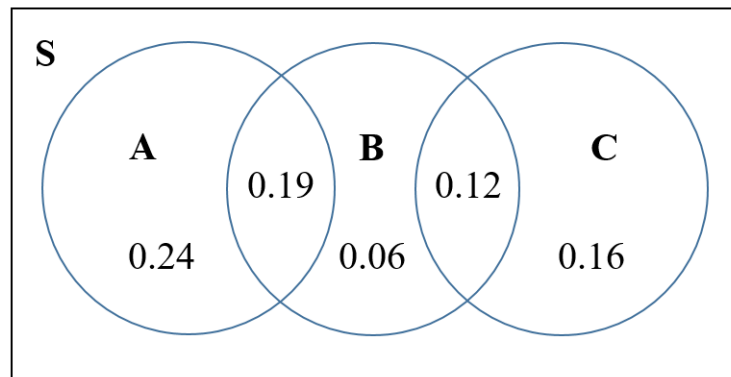
- (c) If two events A and B are independent, we know $P(A \cap B) = P(A)P(B)$. We check for each of the three pairs of events:

$$P(A)P(B) = (0.43)(0.37) = 0.1591 \neq P(A \cap B) \Rightarrow A \text{ and } B \text{ are **not** independent.}$$

$$P(A)P(C) = (0.43)(0.28) = 0.1204 \neq P(A \cap C) \Rightarrow A \text{ and } C \text{ are **not** independent.}$$

$$P(B)P(C) = (0.37)(0.28) = 0.1036 \neq P(B \cap C) \Rightarrow B \text{ and } C \text{ are **not** independent.}$$

- (d) We draw a Venn diagram and fill in each mutually exclusive part, starting with the intersections:



Note that we draw the Venn diagram differently than we usually do. In our regular Venn diagrams for three events, we allow all three circles to overlap. Since we know in this case that $P(A \cap C) = 0$, these two circles don't overlap.

The probability of exactly one event occurring is $P(\text{only } A) + P(\text{only } B) + P(\text{only } C)$. We calculate these probabilities as follows:

$$P(\text{only } A) = P(A) - P(A \cap B) = 0.43 - 0.19 = 0.24.$$

$$P(\text{only } B) = P(B) - P(A \cap B) - P(B \cap C) = 0.37 - 0.19 - 0.12 = 0.06.$$

$$P(\text{only } C) = P(C) - P(B \cap C) = 0.28 - 0.12 = 0.16$$

The probability of exactly one event occurring is $0.24 + 0.06 + 0.16 = 0.46$.

Note that we could also calculate the probability of none of the three events occurring. This corresponds to the area outside the three circles:

$$P(\text{none}) = 1 - (0.24 + 0.19 + 0.06 + 0.12 + 0.16) = 1 - 0.77 = 0.23$$

3. (a) We need to find the critical value z^* for a 93% confidence interval. Draw the standard normal curve for Z , centred at $\mu = 0$. Now draw lines at the values $-z^*$ and z^* on either side of the mean (equally far in either direction), such that the area between $-z^*$ and z^* is 0.93. In order to find the value z^* , we need to find the total area to the left of z^* (so we can use the table to look it up). From the picture, you see that this is equal to the area to the left of $-z^*$ plus the area between $-z^*$ and z^* . We already know $P(-z^* < Z < z^*) = 0.9300$, so we need to find $P(Z < -z^*)$. We know the central area (between $-z^*$ and z^*) is 0.9300, so the total area of the two tails (i.e., below $-z^*$ and above z^*) is $1 - 0.9300 = 0.0700$. That means each tail has an area equal to half this value, $0.0700/2 = 0.0350$. Therefore, the total area to the left of z^* is

$$P(Z < z^*) = P(Z < -z^*) + P(-z^* < Z < z^*) = 0.0350 + 0.9300 = 0.9650.$$

Using the normal table (going backwards), we find that the value of z^* with 0.9650 area to its left is $z^* = 1.81$. (We take the closest value in the table.)

A 93% confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 1000 \pm 1.81 \left(\frac{200}{\sqrt{30}} \right) = 1000 \pm 66.1 = (933.9, 1066.1)$$

- (b) If we took repeated samples of size 30 and calculated the interval in a similar manner, 93% of all such intervals would contain the true mean size of all two-bedroom apartments in Winnipeg.

4. Let $\alpha = 0.05$.

We are testing the hypotheses $H_0: \mu = 150$ vs. $H_a: \mu \neq 150$.

We will reject H_0 if the P-value $\leq \alpha = 0.05$.

$$\text{The test statistic is } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{156 - 150}{14/\sqrt{30}} = 2.35.$$

Note that the test statistic is t rather than z , as we don't know the value of the population standard deviation σ . We only know the sample standard deviation s .

The P-value is $P(T(29) \geq 2.35)$. We see from Table 2 that

$$P(T(29) \geq 2.150) = 0.02 \quad \text{and} \quad P(T(29) \geq 2.462) = 0.01$$

Since $2.150 < t = 2.35 < 2.462$, it follows that $P(T(29) \geq 2.35)$ is between 0.01 and 0.02. But the P-value is double this probability, so the P-value is between 0.02 and 0.04.

Since the P-value $< \alpha = 0.05$, we reject H_0 . At the 5% level of significance, we have sufficient evidence that the true mean weight of all apples differs from 150 grams.

5. (a) We calculate $\hat{p} = \frac{80}{400} = 0.20$. A 90% confidence interval for p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.20 \pm 1.645 \sqrt{\frac{0.20(0.80)}{400}} = 0.20 \pm 0.0329 = (0.1671, 0.2329)$$

- (b) Let $\alpha = 0.10$.

We are testing the hypotheses $H_0: p = 0.25$ vs. $H_a: p < 0.25$.

The test statistic is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.20 - 0.25}{\sqrt{\frac{0.25(0.75)}{400}}} = -2.31$.

The P-value is $P(Z \leq -2.31) = 0.0104$.

Since the P-value $< \alpha = 0.10$, we reject H_0 . At the 10% level of significance, we have sufficient evidence that the true proportion of Canadians who support Donald Trump is less than 0.25.

Sample Final Exam 4 – Part A

Question 1

We use the subscript B to denote the boys, G to denote the girls and C to denote the whole class. The mean height of all students in the class is

$$\bar{x}_C = \frac{n_B \bar{x}_B + n_G \bar{x}_G}{n_B + n_G} = \frac{14(135.2) + 9\bar{x}_G}{23} = \frac{1892.8 + 9\bar{x}_G}{23} = 132.5$$

$$\Rightarrow \bar{x}_G = \frac{23(132.5) - 1892.8}{9} = \frac{1154.7}{9} = 128.3$$

Question 2

We calculate the fences:

$$LF = Q1 - 1.5IQR = 0.256 - 1.5(0.288 - 0.256) = 0.208$$

$$UF = Q3 + 1.5IQR = 0.288 + 1.5(0.288 - 0.256) = 0.336$$

There are two values below the lower fence (0.178 and 0.202) and two values above the upper fence (0.338 and 0.340). These four values would be labelled as outliers. The whisker on the left side of the box extends to the lowest value that is not an outlier, which is 0.210 (i.e., the “new minimum”). The whisker on the right side of the box extends to the highest value that is not an outlier, which is 0.327 (i.e., the “new maximum”).

Question 3

We are using the number of drinks consumed to predict a person’s blood alcohol level, so number of drinks is the explanatory variable X and blood alcohol level is the response variable Y . The fraction of variation in Y explained by its regression on X is, by definition, equal to r^2 . Here, $r^2 = (0.88)^2 = 0.77$, so 77% of the variation in blood alcohol level can be accounted for by its regression on number of alcoholic beverages consumed.

Question 4

There are two factors in this experiment – brand of sunscreen and SPF level.

Brand of sunscreen has two factor levels – Coppertone and Ombrelle. SPF level has two factor levels – 30 and 60.

As such, there are $2 \times 2 = 4$ treatments – Coppertone/SPF 30, Coppertone/SPF 60, Ombrelle/SPF 30 and Ombrelle/SPF 60.

Question 5

Let x be the minimum grade required to get a C. We know that $P(X < x) = 0.10 + 0.08 = 0.18$, since 10% of students get a D and 8% of students get an F. We first find the value z such that $P(Z < z) = 0.1800$ to be $z = -0.92$. (We take the closest value in the table.) Now we use the z -score formula to solve for x as follows:

$$z = \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma = 67 + (-0.92)(11) = 56.88$$

Question 6

We want to find the value of μ such that $P(X < 150) = 0.0700$. We first find the value z such that $P(Z < z) = 0.0700$ to be $z = -1.48$. Now we use the z -score formula to solve for μ as follows:

$$z = \frac{x - \mu}{\sigma} \Rightarrow \mu = x - z\sigma = 150 - (-1.48)(4) = 155.92$$

Question 7

We know that $P(B \cup G) = P(B) + P(G) - P(B \cap G)$, and so

$$P(B) = P(B \cup G) - P(G) + P(B \cap G) = 0.50 - 0.20 + 0.10 = 0.40$$

Question 8

We know none of the pairs of events are mutually exclusive, since $P(J \cap B) > 0$, $P(J \cap G) > 0$ and $P(B \cap G) > 0$. We know if two events A and B are independent, then $P(A \cap B) = P(A)P(B)$. We check this for each pair of events:

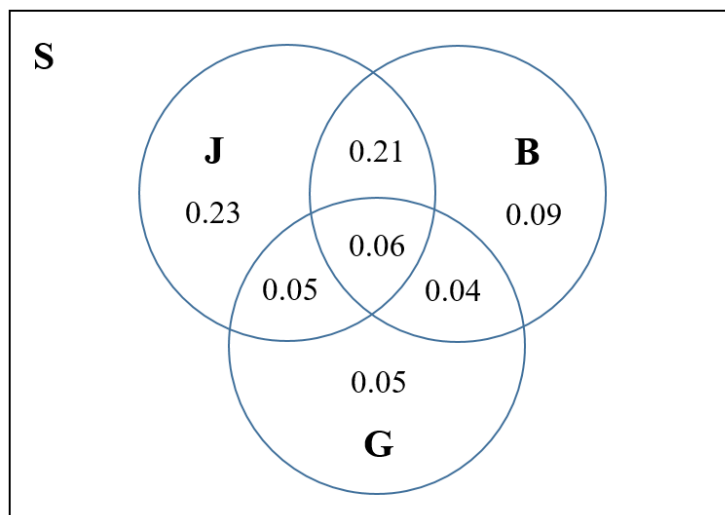
$$P(J)P(B) = (0.55)(0.40) = 0.22 \neq P(J \cap B), \text{ so } J \text{ and } B \text{ are **not** independent.}$$

$$P(J)P(G) = (0.55)(0.20) = 0.11 = P(J \cap G), \text{ so } J \text{ and } G \text{ **are** independent.}$$

$$P(B)P(G) = (0.40)(0.20) = 0.08 \neq P(B \cap G), \text{ so } B \text{ and } G \text{ are **not** independent.}$$

Question 9

We draw a Venn diagram, finding the probability of each mutually exclusive part (starting from the inside with $P(J \cap B \cap G)$ and working our way out). The Venn diagram is shown below:



By adding all probabilities inside the three circles, we find the probability a Winnipegger is a fan of at least one of the three teams is 0.73. It follows that the probability a person is a fan of none of the three teams is

$$P(J^c \cap B^c \cap G^c) = 1 - P(J \cup B \cup G) = 1 - 0.73 = 0.27$$

Question 10

Let X be the number of the 12 people that are Goldeyes fans. There is a fixed number of observations ($n = 12$) and there are only two outcomes of interest for each person (either they are a Goldeyes fan or they aren't). Whether one person is a Goldeyes fan is independent of any other person, and there is a constant probability of success ($p = 0.20$). Therefore,

$$P(X = 3) = \binom{12}{3} (0.20)^3 (0.80)^9 = 0.2362$$

Question 11

Let X be the number of the 500 people that are Jets fans. Then X has a binomial distribution with parameters $n = 500$ and $p = 0.55$. Since the sample size is very large, we can use the normal approximation (you can check that np and $n(1 - p)$ are both greater than 10):

$$P(\hat{p} < 0.50) \approx P\left(Z < \frac{0.50 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(Z \geq \frac{0.50 - 0.55}{\sqrt{\frac{0.55(0.45)}{500}}}\right) = P(Z < 2.25) = 0.0122$$

Question 12

The probability of selecting exactly one yellow ball is

$$\begin{aligned} &P(YY^cY^cY^c) + P(Y^cYY^cY^c) + P(Y^cY^cYY^c) + P(Y^cY^cY^cY) \\ &= (0.1)(0.8)(0.7)(0.6) + (0.9)(0.2)(0.7)(0.6) + (0.9)(0.8)(0.3)(0.6) + (0.9)(0.8)(0.7)(0.4) \\ &= 0.0336 + 0.0756 + 0.1296 + 0.2016 = 0.4404 \end{aligned}$$

Note that we can multiply probabilities for the different boxes, since the colour selected from one box is independent of any other. Note also that when finding the probability of *not* selecting a yellow ball from a box, we do not need to distinguish between red and green. We simply find $P(Y^c)$ as $1 - P(Y)$.

Question 13

The probability of selecting at least one yellow is

$$\begin{aligned} P(\text{at least one yellow}) &= 1 - P(\text{no yellow}) = 1 - P(Y^cY^cY^cY^c) = 1 - (0.9)(0.8)(0.7)(0.6) \\ &= 1 - 0.3024 = 0.6976 \approx 0.7 \end{aligned}$$

Question 14

Consider Case (I). There is a fixed number of trials ($n = 3$) and there are only two outcomes of interest for each selection (either the ball is red or it isn't). Since selections are made with replacement, trials are independent with a constant probability of success ($p = 0.4$). Therefore, X has a binomial distribution with parameters $n = 3$ and $p = 0.4$.

Consider Case (II). The probability of selecting a yellow ball is not constant for all trials. Therefore, X does **not** have a binomial distribution.

Consider Case (III). Since selections are made without replacement, trials are not independent, and the probability of success is not constant. Therefore, X does **not** have a binomial distribution.

Consider Case (IV). There is a fixed number of trials ($n = 4$) and there are only two outcomes of interest for each selection (either the ball is green or it isn't). Trials are independent with constant probability of success ($p = 0.5$). Therefore, X has a binomial distribution with parameters $n = 4$ and $p = 0.5$.

Question 15

Note that we are taking a large sample from a population and plotting the INDIVIDUAL VALUES. This question has nothing to do with a normal distribution or the Central Limit Theorem. (We are NOT plotting sample means.) All we are doing is taking a large sample from the population and plotting the values, so we expect to get something similar to the original population distribution. Think about it the other way around – if we had no idea what the population looked like, and we took a sample of 400 values and plotted them....then if we saw a distribution that looked like a rectangle, it would be logical to conclude the population these values came from is uniform.

Question 16

This one IS about the Central Limit Theorem. Although the population is not normally distributed, we are taking a large sample size, so the sampling distribution of $\bar{X}$ is approximately normal. The mean and standard deviation of $\bar{X}$ are $\mu = 3.5$ and $\sigma/\sqrt{n} = 2/\sqrt{400} = 0.1$.

Question 17

$$P(\bar{X} > 780) = P\left(Z > \frac{780 - 800}{200/\sqrt{100}}\right) = P(Z > -1.00) = 1 - 0.1587 = 0.8413$$

Question 18

$$\begin{aligned}
P(\text{Total} < 30,000) &= P\left(\bar{X} < \frac{30,000}{40}\right) = P(\bar{X} < 750) = P\left(Z < \frac{750 - 800}{200/\sqrt{40}}\right) \\
&= P(Z < -1.58) = 0.0571
\end{aligned}$$

Question 19

Note that we are told that the VARIANCE of X is 144. This is σ^2 , so the standard deviation of X is $\sqrt{144} = 12$. So we know $X \sim N(43, 12)$.

The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$, which we know is equal to 0.8, so we solve for n as follows:

$$\frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{n}} = 0.8 \quad \Rightarrow \quad \sqrt{n} = \frac{12}{0.8} = 15 \quad \Rightarrow \quad n = (15)^2 = 225$$

Question 20

The variability of a sample proportion is measured by its standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

We take a sample of 3% of students at each university, so the sample sizes for U of W and U of M are, respectively, $n_W = 0.03(9000) = 270$ and $n_M = 0.03(27000) = 810$. We are told that $p = 0.2$ for both universities.

The standard deviation of $\hat{p}$ for U of W is $\sqrt{\frac{0.2(0.8)}{270}} = 0.0243$.

The standard deviation of $\hat{p}$ for U of M is $\sqrt{\frac{0.2(0.8)}{810}} = 0.0141$.

The sample proportion for U of W has more sampling variability than that for U of M.

Some people choose (C), because they see that the same percentage of students are selected from each university. However, the variability is only equal if we take the same SAMPLE SIZE from each university – here we are taking the same PERCENTAGE (which results in a higher sample size from U of M). It is true that the population size doesn't affect the variability, but this is if we take the SAME sample size from both populations.

Question 21

We need to find the critical value z^* for a 94% confidence interval. Draw the standard normal curve for Z , centred at $\mu = 0$. Now draw lines at the values $-z^*$ and z^* on either side of the mean (equally far in either direction), such that the area between $-z^*$ and z^* is 0.94. In order to find the value z^* , we need to find the total area to the left of z^* (so we can use the table to look it up). From the picture, you see that this is equal to the area to the left of $-z^*$ plus the area between $-z^*$ and z^* . We already know $P(-z^* < Z < z^*) = 0.9400$, so we need to find $P(Z < -z^*)$. We know the central area (between $-z^*$ and z^*) is 0.9400, so the total area of the two tails (i.e., below $-z^*$ and above z^*) is $1 - 0.9400 = 0.0600$. That means each tail has an area equal to half this value, $0.0600/2 = 0.0300$. Therefore, the total area to the left of z^* is

$$P(Z < z^*) = P(Z < -z^*) + P(-z^* < Z < z^*) = 0.0300 + 0.9400 = 0.9700.$$

Using the normal table (going backwards), we find that the value of z^* with 0.9700 area to its left is $z^* = 1.88$. (We take the closest value in the table.) The required sample size is therefore

$$n = \left(\frac{z^* \sigma}{m} \right)^2 = \left(\frac{1.88(30)}{5} \right)^2 = 127.2384 \approx 128$$

Remember, in sample size calculations, we always round UP.

Question 22

The formula for the confidence interval is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

We know that $\bar{x}$ is at the centre of the confidence interval, so $\bar{x}$ is the average of the lower and upper confidence limits, i.e.,

$$\bar{x} = \frac{1771.6 + 1928.4}{2} = 1850$$

We also know that $\sigma = 280$ and $z^* = 1.960$ (since this is a 95% confidence interval). We can find z^* using either the lower or upper limit of the confidence interval. We will use the upper limit:

$$\bar{x} + z^* \frac{\sigma}{\sqrt{n}} = 1928.4$$

$$\Rightarrow 1850 + 1.960 \left(\frac{280}{\sqrt{n}} \right) = 1928.4$$

$$\Rightarrow \sqrt{n} = \frac{1.960(280)}{(1928.4 - 1850)} = 7 \quad \Rightarrow \quad n = 7^2 = 49$$

Question 23

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{108 - 110}{17/\sqrt{30}} = -0.64.$

The P-value is $P(Z \leq -0.64) = 0.2611.$

Question 24

Statements (A), (B), (D) and (E) are true. Statement (C) is false. Normality is required when we use either z OR t .

Question 25

A 90% confidence interval for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 168 \pm 2.132 \left(\frac{17}{\sqrt{5}} \right) = 168 \pm 16.21 = (155.79, 184.21)$$

Where $t^* = 2.132$ is the upper 0.05 critical value for the t distribution with $n - 1 = 4$ degrees of freedom. Note that we use t rather than z , as the value of the population standard deviation σ is not given.

Question 26

We are testing the hypotheses $H_0: \mu = 150$ vs. $H_a: \mu \neq 150$.

The test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 170 - 150 / 16 / \sqrt{12} = 4.33$.

The P-value is $2P(T(11) \geq 4.33)$. From Table 2, we see that

$$P(T(11) \geq 4.025) = 0.001 \quad \text{and} \quad P(T(11) \geq 4.437) = 0.0005$$

Since $4.025 < t = 4.33 < 4.437$, it follows that $P(T(11) \geq 4.33)$ is between 0.0005 and 0.001. But the P-value is double this probability, so the P-value is between 0.001 and 0.002.

Question 27

The matched pairs t procedures are used when the observations come in pairs – i.e., when observations are dependent (related).

Consider Case (I). We have a pair of observations for each person – their time without hypnosis and their time after hypnosis. Times for each individual are dependent (i.e., related). This is a matched pairs study.

Consider Case (II). We have two independent samples of gas stations – one for Shell and one for Petro Canada. Consider one Shell station in the sample. The gas price at this station is not related to the gas price at any one Petro Canada gas station more than any other. The samples are independent (not related). We make one overall comparison between the two samples. This is **not** a matched pairs study.

Consider Case (III). We have a pair of observations for each married couple – one for the husband and one for the wife. The two observations are dependent (i.e., related). We make a separate comparison for each couple. This is a matched pairs study.

Question 28

The question says “the results are accurate to within 3.1%, 19 times out of 20”. The 3.1% is the margin of error, so $m = 0.031$. The “19 times out of 20” means the confidence level is 95% ($19/20 = 0.95$), so $z^* = 1.960$. The sample size that was used is therefore

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*) = \left(\frac{1.960}{0.031}\right)^2 (0.5)(0.5) = 999.38 \approx 1000$$

Remember, in sample size calculations, we always round UP.

Question 29

We know that $\hat{p}$ must be at the centre of the interval. Therefore, $\hat{p} = (0.33 + 0.37)/2 = 0.35$. Options (A), (B) and (D) are not centered at 0.35, so they are eliminated. Now we are choosing between Options (C) and (E). The interval in (C) is narrower than our 80% confidence interval and the interval in (E) is wider. We know that a 99% confidence interval must be wider than an 80% confidence interval, so that leaves (E) as the correct choice.

Question 30

We are testing the hypotheses $H_0: p = 0.50$ vs. $H_a: p \neq 0.50$. We calculate $\hat{p} = \frac{165}{300} = 0.55$.

The test statistic is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.55 - 0.50}{\sqrt{\frac{0.50(0.50)}{300}}} = 1.73$.

The P-value is $2P(Z \geq 1.73) = 2(1 - 0.9582) = 2(0.0418) = 0.0836$.

Since the P-value $> \alpha = 0.05$, we fail to reject H_0 .

Sample Final Exam 4 – Part B

1. (a) The complete sample space of outcomes is shown below, as well as the probability of each outcome:

Outcome	Probability
GG	$(0.5)(0.6) = 0.30$
GY	$(0.5)(0.1) = 0.05$
GR	$(0.5)(0.3) = 0.15$
YG	$(0.1)(0.6) = 0.06$
YY	$(0.1)(0.1) = 0.01$
YR	$(0.1)(0.3) = 0.03$
RG	$(0.4)(0.6) = 0.24$
RY	$(0.4)(0.1) = 0.04$
RR	$(0.4)(0.3) = 0.12$

Note that we can multiply probabilities for the two sets of lights, as they operate independently.

(b) $P(\text{same colour}) = P(GG) + P(YY) + P(RR) = 0.30 + 0.01 + 0.12 = 0.43$

(c) $P(G_1 \cup R_2) = P(G_1) + P(R_2) - P(G_1 \cap R_2) = 0.5 + 0.3 - (0.5)(0.3) = 0.65$

Note that we could also find this probability by adding the probabilities of all outcomes for which the first light is green or the second light is red (or both).

- (d) Let X be the number of red lights encountered by the student. The variable X can take the values 0, 1 or 2. We find the probability distribution of X as follows:

$$P(X = 0) = P(GG) + P(GY) + P(YG) + P(YY) = 0.30 + 0.05 + 0.06 + 0.01 = 0.42$$

$$P(X = 1) = P(GR) + P(YR) + P(RG) + P(RY) = 0.15 + 0.03 + 0.24 + 0.04 = 0.46$$

$$P(X = 2) = P(RR) = 0.12$$

The probability distribution of X is summarized in the table below:

x	0	1	2
$P(X = x)$	0.42	0.46	0.12

2. (a) A binomial experiment is one for which the following four conditions are satisfied:

- i. There is a fixed number of trials n .
- ii. There are only two possible outcomes of interest for each trial.
- iii. Trials are independent.
- iv. The probability of success p is constant for each trial.

(b) $P(X > 200) = P\left(Z > \frac{200 - 196}{0.8}\right) = P(Z > 0.50) = 1 - 0.6915 = 0.3085.$

(c) Let X be the number of players in the sample who are taller than 200 cm. There is a fixed number of trials ($n = 10$) and there are only two outcomes of interest for each player (either he is taller than 200 cm or he isn't). Whether one player is taller than 200 cm is independent of any other, and there is a constant probability of success ($p = 0.3015$, as calculated in part (b)). Therefore, X has a binomial distribution with parameters $n = 10$ and $p = 0.3085$, and so

$$P(X = 4) = \binom{10}{4} (0.3085)^4 (0.6915)^6 = 0.2080$$

3. (a) A lurking variable is one that explains the relationship between variables in a study, but is not included in the study itself.

(b) A simple random sample of size n is one that is chosen in such a way that all possible groups of n individuals have the same chance of being chosen.

(c) A statistic is a number that describes a sample. Examples of statistics include the sample mean $\bar{x}$, the sample standard deviation s and a sample proportion $\hat{p}$. A parameter is a number that describes an entire population. Examples of parameters include the population mean μ , the population standard deviation σ and a population proportion p .

(d) Two events that are mutually exclusive have no outcomes in common (i.e., they cannot both occur at the same time). Two events that are independent can both occur at the same time, but the outcome of one does not affect the outcome of the other.

(e) Statistical significance is a difference so large that it would rarely be observed by chance alone. We say that the results of a hypothesis test are statistically significant when the P-value $\leq \alpha$ and we reject H_0 .

4. (a) From a histogram of the data, we see the distribution of class sizes for this sample is skewed to the right.
- (b) Although it does not appear that the population distribution of class sizes is normally distributed, the Central Limit Theorem tells us that since the sample size is high (we use $n \geq 30$), the sampling distribution of the sample mean $\bar{X}$ is approximately normal.
- (c) A 99% confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 33.73 \pm 2.576 \left(\frac{18.50}{\sqrt{30}} \right) = 33.73 \pm 8.70 = (25.03, 42.43)$$

- (d) If we took repeated samples of 30 classes and calculated the interval in a similar manner, 99% of such intervals would contain the true mean class size of all third-year classes at the university.
- (e) Let $\alpha = 0.01$.

We are testing the hypotheses $H_0: \mu = 40$ vs. $H_a: \mu \neq 40$.

We will reject H_0 if the P-value $\leq \alpha = 0.01$.

The test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{33.73 - 40}{18.50/\sqrt{30}} = -1.86$.

The P-value is $2P(Z \leq -1.86) = 2(0.0314) = 0.0628$.

Since the P-value $> \alpha = 0.01$, we fail to reject H_0 . At the 1% level of significance, we have insufficient evidence that the true mean size of all third-year classes differs from 40.

5. (a) This is a matched pairs t test. Note that $n = 8$ (and NOT 16), since there are 8 pairs of individuals.

- (b) Let $\alpha = 0.01$.

We are testing $H_0: \mu_d = 0$ vs. $H_a: \mu_d < 0$.

Note that this is a lower-tailed test because we wanted to know if women received FASTER service on average. A fast service time is a LOW service time. Since the differences are defined as Female – Male, if females receive faster service, we would expect the average difference to be negative.

We will reject H_0 if the P-value $\leq \alpha = 0.01$.

The test statistic is $t = \frac{\bar{x}_d}{s_d/\sqrt{n}} = \frac{-27}{29/\sqrt{8}} = -2.63$.

The P-value is $P(T(7) \leq -2.63) = P(T(7) \geq 2.63)$ by the symmetry of the t distributions. From Table 2, we see that

$$P(T(7) \geq 2.517) = 0.02 \quad \text{and} \quad P(T(7) \geq 2.998) = 0.01$$

Since $2.517 < 2.63 < 2.998$, the P-value is between 0.01 and 0.02.

Since the P-value $> \alpha = 0.01$, we fail to reject H_0 . At the 1% level of significance, we have insufficient evidence that females receive faster service on average than males.

- (c) If the service time in nightclubs was the same on average for males and females, the probability of observing a sample mean difference at least as low as -27 would be between 0.01 and 0.02.